Jiayi Zhou

Professor Nicole Dalzell

STA 381

9 December 2022

Write Up on Goal Two Variable Creation

**Section1: Goal**

This research project studies the variations in criminal trial rates on jury trials in felony cases for counties in North Carolina. There are two stages of the project. The second stage is fitting models with variables to explain what causes the variations in criminal trial rates among counties in North Carolina. The first step is collecting explanatory variables. This write-up is about collecting explanatory variables *violent crime rates, property crime rate, the total county population, the percentage of the total population that is Black,* and *the percentage of the total population that is Hispanic ethnicity.*

**Section 2: Data**

For variables: *violent crime rates* and *property crime rate,* I downloaded the *Crime in North Carolina Annual Summaries reports* from 2010 to 2021 from the NC State Bureau of Investigation. The source files are stored in the folder CrimeRatesSourceFiles under Goal2VariableOrganization. The folder OriginalFiles contains the pdf and document format of files. The folder XLSX format has the dataset extracted from table 6 in the reports. It is named based on its year.

For variables: *the total county population, the percentage of the total population that is Black,* and *the percentage of the total population that is Hispanic ethnicity*, we downloaded the dataset from the United States Census Bureau website. The source files are stored in the EthnicMakeupSourceFile under Goal2VariableOrganization. They are named 2010Census and 2020Census. Each of them contains the county population, the population of Hispanics or Latinos, and the population of Black and African Americans for each county in the year of the census.

**Section 3 Data Cleaning, Processing, and Creation**

For explanatory variables– *violent crime rates* and *property crime rates*, I downloaded the reports in pdf format. I transformed them into word documents using Adobe Acrobat. Then I extracted the crime rate data from table 6 in the reports by copying them from the document and pasting them to excel. The datasets are inconsistent in formatting as county names contain subscripts, cells are merged for some rows, and county names are missing for some. Also, I found that for each year, *violent crime rates* and *property crime rates* are listed as different values in two separate reports. So, based on the client's guidance, we keep the rates reported in the later report. Before loading the datasets into R Studio, I manually checked and formatted them. After loading the datasets in R Studio, I added the name of the county for missing values and extracted *violent crime rates* and *property crime rates* for each county from 2010 to 2021. Then I combined all datasets we extracted into one data frame and removed all missing values for later modeling purposes. Some county names contain subscripts which include numbers and commas, so I renamed them to be consistent with the rest of the data we plan to add in the modeling stage. The final data frame is named *crimedata*, with 1008 observations and four variables.

Then I moved on to the find dataset for variables *the total county population, the percentage of the total population that is Black,* and *the percentage of the total population that is Hispanic ethnicity.* From the United States Census Bureau website, other than the decennial census, the only dataset that contains the information on *the total county population, the percentage of the total population that is Black,* and *the percentage of the total population that is Hispanic ethnicity* is the *American Community Survey*. So I downloaded the datasets from 2010 to 2021. After downloading the data, I removed all the explanation sheets in the datasets to load the actual data frame into R Studio. The datasets also contain lots of information we don't need, so I selected the variables we want for each year. After processing the dataset and checking the original file, I found that only around one-third of counties are covered in the dataset.

After discussing with our clients, we used the decennial census in 2010 and 2020. They both have the data for all counties. We downloaded the dataset, removed the explanation page, and loaded them into R Studio. I selected the variables we wanted and named them *Total, HispanicLatino, Black, and County*. I modified the formatting of variables to align with the rest of the datasets we already made, changed the type of variables to numeric, calculated the difference between the years 2020 and 2010, and divided it by 10 to get the static rate of changes. Then, I created estimations for missing years based on the static rate of changes. After that, I combined all the years into one census data and merged it with other explanatory variables for the modeling stage. Before modeling, my teammates and I used correlation plots to explore correlations between variables. We discovered that *Total, HispanicLatino,* and *Black* are highly correlated. Thus, I divided the *Total* by 1000 to make the numeric value smaller, *HispanicLatino* by *Total* to get *the percentage of the total population that is Hispanic ethnicity*, and *Black* by *Total* to get *the percentage of the total population that is Black*. The final combined dataset with all explanatory variables for Y1 is named *Combine2*, with 909 observations and 15 variables. The final combined dataset with all explanatory variables for Y2 is named *Combine3*, with 909 observations and 15 variables.